Lecture notes on stochastics
Athens 2014

A brief introduction to probability

Demetris Koutsoyiannis
Department of Water Resources and Environmental Engineering
School of Civil Engineering
National Technical University of Athens

Presentation available online: http://itia.ntua.gr/1427/

The meaning of probability (by examples)

(1) A fair coin has a probability of 0.5 of heads, and likewise 0.5 of tails; so the probability of tossing two heads in a row is 0.25.

(2) There is a 10% probability of rain tomorrow.

(3) There is a 10% probability of rain tomorrow according to the weather forecast.

(4) Fortunately there is only a 5% probability that her tumor is malignant, but this will not be known for certain until the surgery is done next week.

(5) Smith has a greater probability of winning the election than does Jones.

(6) I believe that there is a 75% probability that she will want to go out for dinner tonight.

(7) I left my umbrella at home today because the forecast called for only a 1% probability of rain.

(8) Among 100 patients in a clinical trial given drug *A,* 83 recovered, whereas among 100 other patients given drug *B,* only 11 recovered; so new patients will have a higher probability of recovery if treated with drug A.

Source of examples: Gauch (2003).

The utility of probability

- Commonly, probability is regarded to be a branch of applied mathematics that provides tools for data analysis.
- Nonetheless, probability is a more general concept that helps shape a consistent, realistic and powerful view of the world.
- Historically, the modern science was initiated from deterministic views of the world, in which probability had a marginal role for peculiar unpredictable phenomena.
- However, in the turn of the nineteenth century, radical developments in physics and mathematics has given the probability theory a central role in the scientific scene, in the understanding and the modelling of natural phenomena.
- Furthermore, probability has provided grounds for philosophical concepts such as indeterminism and causality, as well as for extending the typical mathematical logic, offering the mathematical foundation of induction.
- In typical scientific and technological applications, probability provides the tools to quantify uncertainty, rationalize decisions under uncertainty, and make predictions of future events under uncertainty, in lieu of unsuccessful deterministic predictions.
- Uncertainty seems to be an intrinsic property of nature, as it can emerge even from pure and simple deterministic dynamics, and cannot been eliminated.

See more details in Koutsoyiannis (2008).

Determinism and indeterminism

- The philosophical proposition of *determinism* is widely accepted in science and is manifested in the idea of a clockwork universe.
  - It comes from the French philosopher and scientist René Descartes (1596-1650).
  - It was perfected by the French mathematician and astronomer Pierre-Simon Laplace (1749-1827). It is expressed in the metaphor of *Laplace's demon*, a hypothetical all-knowing entity that knows the precise location and momentum of every atom in the universe at present.
  - The idea encapsulated in the demon metaphor is that, knowing the present perfectly, one can deduce the future and the past using Newton's laws. Thus, according to deterministic thinking, the roots of uncertainty about future are subjective, i.e. rely on the fact that we do not know exactly the present, or we do not have good enough methods and models. It is then a matter of time to eliminate uncertainty, acquiring better data (observations) and building better models.
  - Note though that Isaac Newton (1643-1727) rejected cartesian thinking and especially the clockwork idea; he was aware of the fragility the world and believed that God had to keep making adjustments all the time to correct the emerging chaos.
- In *indeterminism*, a philosophical belief contradictory to determinism, uncertainty may be a structural element of nature and thus cannot be eliminated.
  - Indeterminism has its origin in the Greek philosophers Heraclitus (*ca*. 535–475 BC) and Epicurus (341–270 BC).
  - Its relationship with modern science was theorized by the Austrian-British philosopher Karl Popper (1902-1994).
  - In science, indeterminism largely relies on the notion of *probability*, which according to Popper is the extension (quantification) of the Aristotelian idea of *potentia* (*dynamis*). Practically, the idea is that several outcomes can be produced by a specified cause, while in deterministic thinking only one outcome is possible (but it may be difficult to predict which one).

Deduction and induction

- In mathematical logic, determinism can be paralleled to the premise that all truth can be revealed by *deductive reasoning* or *deduction* (the Aristotelian *apodeixis*). This type of reasoning consists of repeated application of strong syllogisms such as:

  If A is true, then B is true;
  A is true;
  Therefore, B is true.

  If A is true, then B is true;
  B is false;
  Therefore, A is false.

- Deduction uses a set of axioms to prove propositions known as theorems, which, given the axioms, are irrefutable, absolutely true statements. It is also irrefutable that deduction is the preferred route to truth; the question is, however, whether or not it has any limits.
- David Hilbert (1862-1943) expressed his belief that there are no limits in his slogan (from his talk in 1930; also inscribed in his tombstone at Göttingen): “*Wir müssen wissen, wir werden wissen*” (“*We must know, we will know*”). His idea, more formally known as *completeness*, is that any mathematical statement could be proved or disproved by deduction from axioms.
- In everyday life, however, we use weaker syllogisms of the type:

  If A is true, then B is true;
  B is true;
  Therefore, A becomes more plausible.

  If A is true, then B is true;
  A is false;
  Therefore, B becomes less plausible.

- The latter type of syllogism is called *induction* (the Aristotelian *epagoge*). It does not offer a proof that a proposition is true or false and may lead to errors. However, it is very useful in decision making, when deduction is not possible.
- An important achievement of probability is that it quantifies (expresses in the form of a number between 0 and 1) the degree of plausibility of a certain proposition or statement. The formal probability framework uses both deduction, for proving theorems, and induction, for inference with incomplete information or data.

From the almighty determinism of the 19th century to the probabilistic world of the 20th century

1. **Statistical physics** used the probabilistic concept of entropy (which is nothing other than a quantified measure of uncertainty defined within the probability theory) to explain fundamental physical laws (most notably the Second Law of Thermodynamics), thus leading to a new understanding of natural behaviours and to powerful predictions of macroscopic phenomena.
2. **Dynamical systems** theory has shown that uncertainty can emerge even from pure, simple and fully known deterministic (chaotic) dynamics, and cannot be eliminated.
3. **Quantum theory** has emphasized the intrinsic character of uncertainty and the necessity of probability in the description of nature.
4. Developments in **mathematical logic**, and particularly **Gödel's incompleteness theorem**, challenged the almightiness of deduction (inference by mathematical proof) thus paving the road to inductive inference.
5. Developments in **numerical mathematics** highlighted the effectiveness of stochastic methods in solving even purely deterministic problems, such as **numerical integration** in high-dimensional spaces (where a Monte Carlo method is more accurate than a classical deterministic method, and thus preferable for numerical integration, in spaces with more than four dimensions) and **global optimization** of non-convex functions (where stochastic techniques, e.g. evolutionary algorithms or simulated annealing, are in effect the only feasible solution in complex problems that involve many local optima).
6. Advances in **evolutionary biology** emphasize the importance of stochasticity (e.g. in selection and mutation procedures and in environmental changes) as a driver of evolution.

Definition of probability

- According to Kolmogorov's (1933) axiomatization, probability theory is based on three fundamental concepts and four axioms.
- The concepts, i.e., the triplet (Ω, Σ, P) called *probability space*, are:
  1. A non-empty set Ω, sometimes called the *basic set*, *sample space* or the *certain event* whose elements ω are known as *outcomes* or *states*.
  2. A set Σ known as *σ-algebra* or *σ-field* whose elements E are subsets of Ω, known as *events*. Ω and $\emptyset$ are both members of Σ, and, in addition, (a) if E is in Σ then the complement $\Omega - E$ is in Σ; (b) the union of countably many sets in Σ is also in Σ.
  3. A function P called *probability* that maps events to real numbers, assigning each event E (member of Σ) a number between 0 and 1.
- The four axioms, which define the properties of P, are:

  I. **Non-negativity**: For any event A, $P(A) \geq 0$.

  II. **Normalization**: $P(\Omega) = 1$.

  III. **Additivity**: For any events A, B with $AB = \emptyset$, $P(A + B) = P(A) + P(B)$.

  IV. **Continuity at zero**: If $A_1 \supset A_2 \supset \ldots \supset A_n \supset \ldots$ is a decreasing sequence of events, with $A_1 A_2 \ldots A_n \ldots = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$.

  [Note: In the case that Σ is finite, axiom IV follows from axioms I-III; in the general case, however, it should be put as an independent axiom.]

The concept of a random variable

- A random variable $\underline{x}$ is a function that maps outcomes to numbers, i.e. quantifies the sample space Ω.
- More formally, a real single-valued function $\underline{x}(\omega)$, defined on the basic set Ω, is called a *random variable* if for each choice of a real number a the set $\{\underline{x} < a\}$ for all ω for which the inequality $\underline{x}(\omega) < \alpha$ holds true, belongs to Σ.
- With the notion of the random variable we can conveniently express events using basic mathematics. In most cases this is done almost automatically. For instance a random variable $\underline{x}$ that takes values 1 to 6 is intuitively assumed when we deal with a die through.
- We must be attentive that a random variable is not a number but a function. Intuitively, we could think of a random variable as an object that represents simultaneously all possible outcomes and only them.
- A particular value that a random variable may take in a random experiment, else known as a *realization* of the variable, is a number.
- We can denote a random variable by an underlined letter, e.g. $\underline{x}$ and its realization with a non-underlined letter x (another convention is to use an upper case letter, e.g. X, for the random variable and a lower case letter, e.g. x, for its realization. In any case, random variables and values thereof two should not be confused.

Probability distribution function

- *Distribution function* is a function of the real variable x defined by

  $$F(x) := P\{\underline{x} \le x\}$$

  where $\underline{x}$ is a random variable.

- The random variable with which this function is associated is not an argument of the function. If there risk of confusion (e.g. there are many random variables), the random variable is usually denoted as a subscript (e.g. $F_{\underline{x}}(x)$). Typically $F(x)$ has a mathematical expression depending on some parameters. The domain of $F(x)$ is not identical to the range of the random variable $\underline{x}$; rather it is always the set of real numbers.

- The distribution function is a non-decreasing function obeying the relationship

  $$0 = F(-\infty) \le F(x) \le F(+\infty) = 1$$

- For its non-decreasing attitude, in the English literature the distribution function is also known as *cumulative distribution function* (cdf) – though "cumulative" is not necessary. In practical applications the distribution function is also known as *non-exceedence probability*. Likewise, the non-increasing function

  $$\bar{F}(x) = P\{\underline{x} > x\} = 1 - F(x)$$

  is known as *exceedence probability* (or survival function, survivor function, tail function).

- The distribution function is always continuous on the right; however, if the basic set Ω is finite or countable, $F(x)$ is discontinuous on the left at all points x_i that correspond to outcomes ω_i, and it is constant between them (staircase-like). Such random variable is called *discrete*. If $F(x)$ is a continuous function, then the random variable is called *continuous*. A *mixed* case is also possible; in this the distribution function has some discontinuities on the left, but is not staircase-like.

- For continuous random variables, the inverse function $F^{-1}(\ )$ of $F(\ )$ exists. Consequently, the equation $u = F(x)$ has a unique solution for x, called *u-quantile* of the variable $\underline{x}$, that is:

  $$x_u = F^{-1}(u)$$

Probability density (or mass) function

- In continuous variables any particular value *x* has zero probability to occur. However, we can still tell which of two outcomes is more probable by examining the ratio of the two probabilities. As this is a $0/0$ expression, having in mind l'Hôpital's rule, we need to examine the ratio of derivatives of probabilities.
- The derivative of the distribution function is called the *probability density function*:

$$f(x) \coloneqq \frac{\mathrm{d}F(x)}{\mathrm{d}x}$$

- The basic properties of $f(x)$ are

$$f(x) \geq 0, \quad \int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

- Obviously, the probability density function does not represent a probability; therefore it can take values higher than 1. Its relationship with probability is described by the following equation:

$$f(x) = \lim_{\Delta x \to 0} \frac{P\{x \leq \underline{x} \leq x + \Delta x\}}{\Delta x}$$

- The distribution function can be calculated from the density function by

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y$$

- In discrete random variables, the density is a sequence of Dirac δ functions. It is thus more convenient to use the so-called *probability mass function* $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, $j = 1,\ldots,w$, where *w* is the number of possible outcomes (which can be infinite).

Some common distributions

Name	**Probability density function**	**Distribution function**
Uniform in [0, 1]	$f(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$	$F(x) = \max(0, \min(x, 1))$
Exponential	$f(x) = \begin{cases} \mathrm{e}^{-x/\mu} / \mu & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$	$F(x) = \begin{cases} 1 - \mathrm{e}^{-x/\mu} & \text{for } x \ge 0 \\ 0 & \text{for } x < 0 \end{cases}$
Normal	$f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$	$F(x) =$ $\dfrac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp\left(-\dfrac{(y-\mu)^2}{2\sigma^2}\right) dy$

Independent and dependent events, conditional probability

- Two events A and B are called *independent* (or *stochastically independent*), if

$$P(AB) = P(A)P(B)$$

- Otherwise A and B are called (*stochastically*) *dependent*.
- The definition can be extended to many events. Thus, the events A_1, A_2, …, are *independent* if for any finite set of distinct indices i_1, i_2, …, i_n:

$$P\left(A_{i_1} A_{i_2} \ldots A_{i_n}\right) = P\left(A_{i_1}\right) P\left(A_{i_2}\right) \ldots P\left(A_{i_n}\right)$$

- The handling of probabilities of independent events is thus easy. However, this is a special case because usually natural events are dependent. In the handling of dependent events the notion of *conditional probability* is vital.
- By definition (Kolmogorov, 1933), conditional probability of the event A given B (i.e. under the condition that the event B has occurred) is the quotient

$$P(A|B) \coloneqq \frac{P(AB)}{P(B)}$$

- Obviously, if $P(B) = 0$, this conditional probability cannot be defined, while for independent A and B, $P(A|B) = P(A)$. It follows that

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

- From this it follows the *Bayes theorem*:

$$P(B|A) = P(B)\frac{P(A|B)}{P(A)}$$

Random number generation

- **Sequence of random numbers** is a sequence of numbers x_i whose every one statistical property is consistent with that of a sample from a sequence of independent identically distributed random variables $\underline{x}_i$ (adapted from Papoulis, 1990).
- **Random number generator** is a device (typically computer algorithm) which generates a sequence of random numbers x_i with given distribution $F(x)$. As most algorithms are purely deterministic, sometimes the numbers are called pseudorandom—but this in not necessary.
- Random number generation is also known as **Monte Carlo** sampling.
- The basis of practically all random generators is the uniform distribution in [0,1]. A typical procedure is the following:
  - ❑ We generate a sequence of integers q_i from the recursive algorithm

    $$q_i = (k\, q_{i-1} + c) \bmod m$$

    where k, c and m are appropriate integers (e.g. $k = 69\,069$, $c = 1$, $m = 2^{32} = 4\,294\,967\,296$ or $k = 7^5 = 16\,807$, $c = 0$, $m = 2^{31} - 1 = 2\,147\,483\,647$; Ripley, 1987, p. 39).
  - ❑ We calculate the sequence of random numbers u_i with uniform distribution in [0,1] by

    $$u_i = q_i \,/\, m$$

- A more recent and better algorithm is the so-called *Mersenne twister* (en.wikipedia.org/wiki/Mersenne_twister). It is available in most languages and software packages. For example, for Excel (which by default includes the function rand) the Mersenne twister algorithm, called NtRand, can be found in www.ntrand.com/download/.
- A direct (but sometimes time demanding) algorithm to produce random numbers x_i from *any* $F(x)$ given random numbers u_i with uniform distribution in [0,1] is provided by:

  $$x_i = F^{-1}(u_i)$$

Exercise 1

Let $\underline{x}$ and $\underline{y}$ represent the outcomes of each of two dice. What is the probability of the following cases?

- $\{\underline{x} < \underline{y}\}$
- $\{\underline{x} < y\}$
- $\{x < \underline{y}\}$
- $\{x < y\}$

Verify the results by Monte Carlo simulations.

Exercise 2

- Assume that in a certain place on earth and a certain period of the year a dry and a wet day are equiprobable and that in the different days the states (wet/dry) are independent. What is the probability that two consecutive days are wet under the following conditions?
  - ❑ Unconditionally.
  - ❑ If we know that the first day is wet.
  - ❑ If we know that the second day is wet.
  - ❑ If we know that one of the two days is wet.
  - ❑ If we know that one of the two days is dry.
- Verify the results by Monte Carlo simulations.
- Plot the distribution function of one day's state (wet/dry) (after introducing an appropriate random variable).
- Assuming that in a wet day the probability density function of the rainfall depth $\underline{x}$ (expressed in mm) is $f(x|\text{wet}) = \mathrm{e}^{-x}$, plot the probability distribution function $F(x)$.

Exercise 3

- Three engineers A, B and C are biding for a 1 000 000 € project and the evaluation committee in order to make the fairest possible selection, decided to throw a die. If the outcome is 1 or 2 the projects goes to A, if it is 3 or 4, then B wins and if it is 5 or 6, then C wins. The dice is cast, but the announcement of the winner is going to be done the next day by the minister.
- Engineer A approaches the chairman of the committee and offers him 1000 € to accept his following request: "I know you are not allowed to tell me who wins; however, two of the three will lose. Therefore, B or C or both will lose. Please tell me just one of these two will lose". The committee member accepts and says that C will lose. Then engineer A offers another 1000 € to swap him with B.
- Prove that the strategy of engineer A is consistent with awareness of probability.
- Compare this strategy with another one, in which engineer A offers the same amount to convince the chairman to re-decide on A and B by tossing a coin.
- Verify your result with Monte Carlo simulation.

Note: A different utterance of this problem is known as the "three prisoners problem" (http://en.wikipedia.org/wiki/Three_Prisoners_problem), which has puzzled many. For example, Ben-Naim, 2008, devotes several pages in his book about entropy (including a whole appendix) to solve this problem. However, its solution can be done in two lines.

Expectation

- For a discrete random variable $\underline{x}$, taking on the values $x_1, x_2, \ldots, x_w$ (where w could be ∞) with probability mass function $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, if $g(\underline{x})$ is an arbitrary function of $\underline{x}$ (so that $g(\underline{x})$ is a random variable per se), we define the *expectation* or *expected value* or *mean* of $g(\underline{x})$ as

$$\mathrm{E}\left[g\left(\underline{x}\right)\right] \coloneqq \textstyle\sum_{j=1}^{w} g\left(x_j\right) P\left(x_j\right)$$

- Likewise, for a continuous random variable $\underline{x}$ with density $f(x)$, the expectation is

$$\mathrm{E}\left[g\left(\underline{x}\right)\right] \coloneqq \textstyle\int_{-\infty}^{\infty} g(x) f(x) \mathrm{d}x$$

- For certain types of functions $g(\underline{x})$ we get very commonly used statistical parameters, as specified below:

1. For $g(\underline{x}) = \underline{x}^r$, where $r = 0, 1, 2, \ldots$, the quantity $\mu_r := \mathrm{E}[\underline{x}^r]$ is called the r*th moment* (or the r*th moment about the origin*) of $\underline{x}$. For $r = 0$, obviously the moment is 1.
2. For $g(\underline{x}) = \underline{x}$, the quantity $\mu := \mu_1 = \mathrm{E}[\underline{x}]$(that is, the first moment) is called the *mean* of $\underline{x}$.
3. For $g(\underline{x}) = (\underline{x} - \mu)^r$ where $r = 0, 1, 2, \ldots$, the quantity $m_r := \mathrm{E}[(\underline{x} - \mu)^r]$ is called the r*th central moment* of $\underline{x}$. For $r = 0$ and 1 the central moments are respectively 1 and 0. For
4. For $g(\underline{x}) = (\underline{x} - \mu)^2$ the quantity $\sigma^2 := m_2 = \mathrm{E}[(\underline{x} - \mu)^2]$ is called the *variance* of $\underline{x}$ (also denoted as var[$\underline{x}$]); its square root σ (also denoted as std[$\underline{x}$] is called the standard deviation of $\underline{x}$.

Entropy

- For a **discrete random variable** $\underline{x}$, taking on the values $x_1, x_2, \ldots, x_w$ (where w could be ∞) with probability mass function $P_j \equiv P(x_j) = P\{\underline{x} = x_j\}$, the *entropy* is defined as the expectation of the minus logarithm of probability (Shannon, 195?), i.e.:

$$\Phi[\underline{z}] := \mathrm{E}[-\ln P(\underline{z})] = -\sum_{j=1}^{w} P_j \ln P_j$$

- Extension of the above definition for the case of a **continuous random variable** $\underline{x}$ with probability density function $f(x)$, is possible, although not contained in Shannon's (1948) original work. This extension involves a (so-called) '*background measure*' with density $h(x)$, which can be any probability density, proper (with integral equal to 1) or improper (meaning that its integral does not converge); typically it is an (improper) Lebesgue density, i.e. a constant with dimensions $[h(x)] = [f(x)] = [x^{-1}]$, so that the argument of the logarithm function that follows be dimensionless. Thus, the entropy of a continuous variable $\underline{x}$ is (see e.g. Jaynes, 2003, p. 375):

$$\Phi[\underline{x}] := \mathrm{E}\left[-\ln \frac{f(\underline{x})}{h(\underline{x})}\right] = -\int_{-\infty}^{\infty} \ln \frac{f(x)}{h(x)} f(x)\,\mathrm{d}x$$

- It is easily seen that for both discrete and continuous variables the entropy $\Phi[\underline{z}]$ is a *dimensionless* quantity.
- The importance of the entropy concepts relies in the **principle of maximum entropy** (Jaynes, 1957); it postulates that the entropy of a random variable $\underline{z}$ should be at maximum, under some conditions, formulated as constraints, which incorporate the information that is given about this variable.
- This principle can be used for **logical inference** as well as for **modelling physical systems**; for example, the tendency of entropy to become maximal (Second Law of thermodynamics) can result from this principle.

Exercise 4

- Find the mean, variance and entropy of the variable $\underline{x}$ representing the outcome of a fair die. Show that the entropy of a fair die is greater than in any loaded die.
- Find the mean, variance and entropy of a variable $\underline{x}$ with uniform distribution in [0,1]. Show that this entropy is the maximum possible among all distributions in [0,1].
- Find the mean, variance and entropy of a variable $\underline{x}$ with exponential distribution. Show that this entropy is the maximum possible among all distributions in $[0,\infty)$ which have specified mean.
- Find the mean, variance and entropy of a variable $\underline{x}$ with normal distribution. Show that this entropy is the maximum possible among all distributions in $(-\infty,\infty)$ which have specified mean and variance.

References

- Ben-Naim, A., *A Farewell to Entropy: Statistical Thermodynamics Based on Information*, World Scientific Pub., Singapore, 384 pp., 2008.
- Gauch, H.G., Jr., *Scientific Method in Practice*, Cambridge University Press, Cambridge, 2003.
- Jaynes, E.T. Information theory and statistical mechanics, *Physical Review*, 106 (4), 620-630, 1957.
- Jaynes, E.T. *Probability Theory: The Logic of Science*, Cambridge Univ. Press, Cambridge, 728 pp., 2003.
- Kolmogorov, A. N., Grundbegrijfe der Wahrscheinlichkeitsrechnung, *Ergebnisse der Math.* (2), Berlin, 1933; 2nd English Edition: Foundations of the Theory of Probability, 84 pp. Chelsea Publishing Company, New York, 1956.
- Koutsoyiannis, D., *Probability and statistics for geophysical processes*, National Technical University of Athens, Athens, 2008 (itia.ntua.gr/1322/).
- Papoulis, A., *Probability and Statistics*, Prentice-Hall, New Jersey, 1990.
- Ripley, B. D., *Stochastic Simulation*, Wiley, New York, 1987.
- Shannon, C.E. The mathematical theory of communication, *Bell System Technical Journal*, 27 (3), 379-423, 1948.